doi:10.1006/geno.2001.6683, available online at http://www.idealibrary.com on IDEAL

Gene Expression Profile of Human Bone Marrow Stromal Cells: High-Throughput Expressed Sequence Tag Sequencing Analysis

Libin Jia,[1] Marian F. Young,[2] John Powell,[3] Liming Yang,[3] Nicola C. Ho,[4] Robert Hotchkiss,[5] Pamela Gehron Robey,[2] and Clair A. Francomano[4,*]

[1]*Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*
[2]*Craniofacial and Skeletal Diseases Branch, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland 20892, USA*
[3]*Bioinformatics and Molecular Analysis Section, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, USA*
[4]*Laboratory of Genetics, National Institute of Aging, National Institutes of Health, Baltimore, Maryland 21224, USA*
[5]*Hospital for Special Surgery, New York, New York 10021, USA*

**To whom correspondence and reprint requests should be addressed. Fax: (301) 558-8087. E-mail: francomanocl@grc.nia.nih.gov.*

Human bone marrow stromal cells (HBMSC) are pluripotent cells with the potential to differentiate into osteoblasts, chondrocytes, myelosupportive stroma, and marrow adipocytes. We used high-throughput DNA sequencing analysis to generate 4258 single-pass sequencing reactions (known as expressed sequence tags, or ESTs) obtained from the 5′ (97) and 3′ (4161) ends of human cDNA clones from a HBMSC cDNA library. Our goal was to obtain tag sequences from the maximum number of possible genes and to deposit them in the publicly accessible database for ESTs (dbEST of the National Center for Biotechnology Information). Comparisons of our EST sequencing data with nonredundant human mRNA and protein databases showed that the ESTs represent 1860 gene clusters. The EST sequencing data analysis showed 60 novel genes found only in this cDNA library after BLAST analysis against 3.0 million ESTs in NCBI's dbEST database. The BLAST search also showed the identified ESTs that have close homology to known genes, which suggests that these may be newly recognized members of known gene families. The gene expression profile of this cell type is revealed by analyzing both the frequency with which a message is encountered and the functional categorization of expressed sequences. Comparing an EST sequence with the human genomic sequence database enables assignment of an EST to a specific chromosomal region (a process called digital gene localization) and often enables immediate partial determination of intron/exon boundaries within the genomic structure. It is expected that high-throughput EST sequencing and data mining analysis will greatly promote our understanding of gene expression in these cells and of growth and development of the skeleton.

Key Words: human bone marrow stromal cells, EST sequencing analysis, novel genes, gene expression profile, data mining

INTRODUCTION

Human bone narrow stromal cells (HBMSC) are nonhematopoietic cells residing in the marrow cavity. They have many characteristics of stem cells for tissues that can roughly be defined as mesenchymal, because they can differentiate into osteoblasts, chondrocytes, myelosupportive stroma, adipocytes, and even myoblasts [1]. Therefore, bone marrow stromal cells present an intriguing model for examining the differentiation of stem cells. Also, several characteristics make them potentially useful for cell and gene therapy [2–4]. After extensive proliferation *in vitro*, the HBMSC population includes precursor cells for at least four types of connective tissue: bone, cartilage, hematopoiesis-supporting stroma, and associated adipocytes [5–8]. When single bone marrow stromal cells develop into individual HBMSC colonies, they show different morphologies and rates of proliferation. HBMSC strains derived from individual colonies also vary widely in their ability to form bone and the hematopoietic microenvironment after *in vivo* transplantation [9]. Despite efforts to understand the biology of HBMSC through studies

0888-7543/01 $35.00

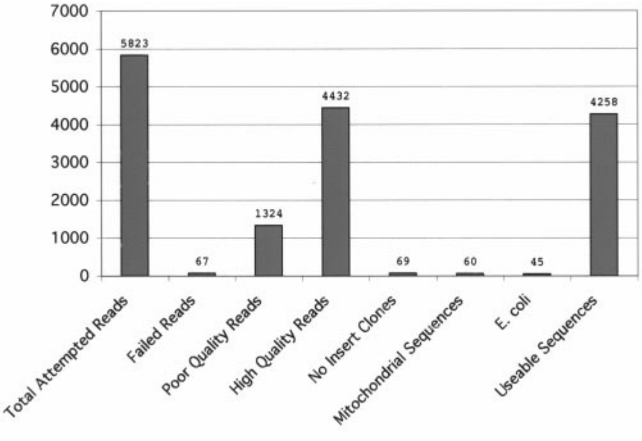

FIG. 1. Evaluation of the HBMSC cDNA library. Parameters 1–4 indicate EST sequencing numbers for obtaining high-quality reads. Parameters 5–7 indicate background EST sequences. Parameter 8 indicates the number of final usable EST sequences.

at different levels, the study of genes and proteins important to the biological phenotypes of HBMSC is still in its infancy. Well-known exceptions include *STRO1* [10,11], *THY1* [12,13], and *SCA1* [14]. Determining the genetic expression profile of this specific cell type is key to rapid advances in understanding skeletal growth and development. The ability to peer into HBMSC and read their molecular signature will enable us to identify more precisely differences in gene expression that make a specific derivative cell type unique. It should also uncover specific and sensitive molecular markers for bone tissue growth and development.

In 1997, The National Human Genome Research Institute (NHGRI) and National Institute of Dental and Craniofacial Research, along with other institutes (Center for Information Technology, National Cancer Institute (NCI), and National Center for Biotechnology Information (NCBI)) at the National Institutes of Health (NIH), and the Hospital for Special Surgery in New York City launched the Human Skeletal Genome Anatomy Project (SGAP) [15]. The overall goal of SGAP is comprehensive molecular characterization of skeletal-related cells and tissues. SGAP is a resource for scientists interested in normal and abnormal skeletal growth and development and was developed in parallel to CGAP, the Cancer Genome Anatomy Project, directed by the NCI [16]. SGAP will include a catalog of genes expressed in bone and cartilage as well as a tissue bank of normal and abnormal bone and cartilage, tendon, ligament, and synovium. It is intended that genes included in SGAP will be those necessary for general bone growth and development, mutations of which result in the skeletal dysplasias and related monogeneic and complex disorders of skeletal growth and development. Establishment of an index of genes expressed in skeletal cells (skeletal gene index) is an essential step in support of the goal of a complete molecular analysis of bone cells [17].

An initial goal of SGAP is to characterize gene expression patterns of specific cell types. Here, we report the results of large-scale, high-throughput sequencing of expressed sequence tags (ESTs) derived from HBMSC.

Large-scale, single-pass sequencing of cDNA clones randomly picked from libraries has proven to be a powerful approach to discovering genes and novel members of gene families as well as an expressed gene profile [18–26]. The NCBI EST database (dbEST) has become one of the fastest growing segments of the public DNA databases [27]. ESTs are DNA sequences read from one or both ends of expressed gene fragments. The Merck-Washington University EST Project [28] and several other public EST projects are rapidly discovering the complement of human genes and making them easily accessible. Although incomplete and not error-free, ESTs remain an effective means for novel gene discovery and generating biologically informative probes for mapping genes to chromosomes as sequence-tagged sites (STSs), for identifying mutations leading to heritable diseases, and for full-length cDNA cloning [21,26]. The advantages of this approach are as follows: (1) it can be pursued as a relatively inexpensive and rapid way to access many of the expressed genes of a cell or tissue type [29,30]; and (2) with the advent of high-throughput sequencing technology and an increased interest in genome-wide studies, it became clear that ESTs could be generated in sufficient numbers to provide a rapid means of gene discovery [28,31], especially for those searching for human disease genes or constructing physical maps of the human genome [32].

Based on the feasibility of EST analysis for gene discovery and pattern configuration in other cell types or organisms [31], we undertook a larger-scale EST project to sequence randomly isolated cDNAs from a HBMSC cDNA library. The

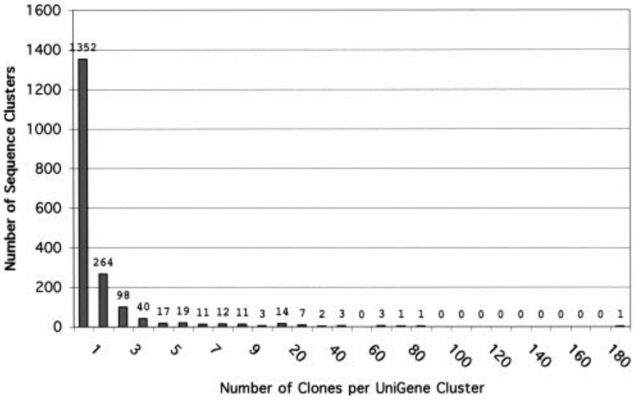

FIG. 2. Sequence clusters in HBMSC cDNA library ESTs. The set of 1860 apparently nonrepetitive ESTs (Fig. 1) were subjected to sequence neighboring, and overlapping sequences were grouped into clusters. The number of unique clusters is indicated as a function of the number of ESTs in each cluster. Thus, 1352 clusters contain only a single EST, whereas 508 clusters contain two or more ESTs.

analysis results from 4258 cDNA sequences demonstrate that large-scale, high-throughput EST sequencing and data analysis are powerful means for identifying novel genes and an expression profile as well as for mapping genes expressed in this cell type.

Results

The analysis presented here was performed on 4258 ESTs generated from the 4258 clones sequenced as of July 1, 1999. We obtained all the sequences from oligo(dT)-primed directionally cloned cDNAs.

General Analysis of HBMSC cDNA Library

We obtained single-pass sequences from randomly selected clones from the HBMSC cDNA library. So far, 5823 ESTs have been sequenced from this library. To assess the various types of background in the library, we compared the 5823 ESTs sequenced at the NIH Intramural Sequencing Center (NISC) and at the Washington University with cloning vector, human or mouse mitochondria, bacterial (*Escherichia coli* genomic), human or mouse ribosomal, *Alu*, and other repeats. The results of these analyses are summarized in Fig. 1. In general, the relative representation of background sequences in this cDNA library is low (< 2% for each type of contamination: human mitochondrial, bacterial, and vector). No ribosomal DNA or mouse mitochondrial DNA was found. However, 1.6% of the clones did not have an insert. Although this is higher than is usually found among libraries constructed by Stratagene, we considered this cDNA library as acceptable for doing high-throughput sequencing analysis. The identified background sequences were separated from our collection, resulting in a final set of 4258 ESTs and an overall successful sequencing rate of 74%. These ESTs have been deposited in NCBI dbEST and are listed on the Web (http://www.ncbi.nlm.nih.gov/UniGene/lib.cgi?ORG=Hs&LID=574).

Sequence Redundancy

To assess the complexity and depth of the direct selection cDNA library, we performed a sequence neighboring analysis on the set of 4258 ESTs from HBMSC. This analysis entails comparing each EST with all others in a pairwise fashion, allowing the sequences to be grouped into clusters [32]. There are a total of 1860 gene sets (Fig. 2). Among them, 1352 ESTs were only found once; that is, a similar sequence was not encountered in our collection. The remaining clones identified one or more sequence neighbors. These sequences formed 508 clusters, with most containing two or three overlapping sequences. Thus, the ESTs reported here form a nonredundant gene set of 1860 sequence clusters. Most sequences occurred only once; 73% were singletons. Highly redundant sequences, those occurring more than five times, made up ~ 6% of all successful reads. In total, 1860 unique sequences were identified by combining all singletons plus the number of distinct clusters with two or more members.

Novel Gene Discovery

One of the major goals for high-throughput EST sequencing is gene discovery. Identification of novel genes (genes that are uniquely expressed in this cell type) is of considerable interest, both in biological terms and as potential targets for drug or vaccine design. We submitted single-pass sequencing from the 5′ (97 clones) or the 3′ (4161 clones) end of 4258 independent PCR-amplified cDNAs from the HBMSC cDNA library to the NCBI GenBank (dbEST). The average length of the high-quality sequences was 468 bp, sufficient to allow robust sequence homology searches. We used these ESTs to search against the nonredundant GenBank (nucleic acid and protein database) and dbEST database of NCBI using BLAST with various search parameters to create listings of novel genes. BLAST homology searches against NCBI's dbEST database and nonredundancy GenBank were performed. We considered an EST (as a gene transcript) novel if there was no hit for the BLAST search for this particular EST. A hit was defined with a sliding identity match percentage cut-off of 96% over a 100-bp window, or of 98% for a > 50-bp window. Table 1 shows that, among the 4258 ESTs, about 60 novel ESTs are found only in this library. The ratio for novel ESTs is about 1.4%. Because most of the HBMSC ESTs were sequenced from the 3′ end, we also searched the poly(A) site, AAATAA (the reverse sequence is TTTATT) or AAATTA (the reverse sequence is TTTAAT), for these novel ESTs. Table 1 lists the novel ESTs found only in this HBMSC cDNA library.

HBMSC Gene Expression Profile

To obtain the gene expression pattern and enhance the biological value of the data, we analyzed the BLAST results using various match stringencies to create listings of putative genes. ESTs were clustered by sequence similarity to reveal the number of times a given sequence was encountered. Table 2 lists the 30 most abundant genes expressed in HBMSC. The majority of the 30 most abundant sequences are previously identified genes that encode a variety of cellular matrix or secretory proteins. The most frequently expressed gene encoded fibronectin (188 times, 4.65%), followed by those encoding type I collagen, α2 (90 times, 2.23%); type I collagen, α1 (82 times, 2.03%); osteonectin (78 times, 1.93%); eukaryotic translation elongation factor 1, α1 (74 times, 1.83%); γ1-actin (71 times, 1.76%); β-actin (66 times, 1.63%); transgelin (44 times, 1.09%), ferritin, heavy chain (42 times, 1.04%); and annexin II (41 times, 1.02%). Other highly expressed genes (over 0.5% of total mRNA) encoded connective tissue growth factor (31 times, 0.77%); transforming growth factor, β-induced, 68 kD (31 times, 0.77%); human normal keratinocyte subtraction library mRNA, clone H22a (29 times, 0.72%); vimentin (28 times, 0.69%); tubulin-α (27 times, 0.67%); human aortic-type smooth muscle α-actin (27 times, 0.67%); insulin-like growth factor-binding protein 4 (25 times, 0.62%); and plasminogen activator inhibitor, type 1 (24 times, 0.59%). Their relative abundance suggests that they encode proteins with important roles in the biology of HBMSC. Indeed, recent studies of osteonectin-null mice show they have decreased bone formation and decreased osteoblast and osteoclast surface and

doi:10.1006/geno.2001.6683, available online at http://www.idealibrary.com on IDEAL

TABLE 1: Novel ESTs found in HBMSC cDNA library

Clone number	Hs.ID	GenBank ID	EST read length (bp)	Seq. read direction	PolyA_Sig position	PolyA_Seq (-)
1	Hs.112513	AA599376	409	3′	16 - 21	TTTATT
2	Hs.112525	AA600076	406	3′	26 - 31	TTTAAT
3	Hs.112532	AA600238	465	3′	19 - 24	TTTATT
4	Hs.116692	AA669885	412	3′	17 - 22	TTTATT
5	Hs.135191	AA600037	218	3′		
6	Hs.139876	AA599924	419	3′	1 - 6	TTTATT
7	Hs.126709	AA703947	490	3′		
8	Hs.140044	AA666148	289	3′	20 - 25	TTTAAT
9	Hs.146299	AA545764	479	5′		
10	Hs.162615	AA600059	423	3′	1 - 6	TTTATT
11	Hs.169522	AA669780	249	3′	20 - 25	TTTATT
12	Hs.188120	AA626922	230	3′		
13	Hs.193157	AA669940	408	3′		
14	Hs.203741	AA664436	446	3′		
15	Hs.204460	AI753538	494	3′	482 - 487	ATTAAA (+)
16	Hs.204584	AI754246	410	3′		
17	Hs.204585	AI754827	509	3′	45 - 50	TTTATT
18	Hs.205786	AI754862	503	3′	50 - 55	TTTAAT
19	Hs.205787	AI754897	119	3′		
20	Hs.204747	AI753729	391	3′	355 - 360	AATAAA(+)
21	Hs.204748	AI753988	461	3′		
22	Hs.204750	AI754625	379	3′	33 - 38	TTTAAT
23	Hs.204930	AI752834	372	3′		
24	Hs.204931	AI754059	510	3′	33 - 38	TTTATT
25	Hs.205427	AI753557	498	3′		
26	Hs.205428	AI753655	456	3′	444 - 449	AATAAA(+)
27	Hs.205429	AI753683	133	3′		
28	Hs.205430	AI754201	497	3′	19 - 24	TTTATT
29	Hs.205342	AI754628	426	3′	51 - 56	TTTATT
30	Hs.205433	AI754789	517	3′	33 - 38	TTTATT
31	Hs.205434	AI755069	293	3′	30 - 35	TTTATT
32	Hs.205781	AI753390	462	3′	17 - 22	TTTATT
33	Hs.205782	AI753813	459	3′	436 - 441	AATAAA(+)
34	Hs.205784	AI754692	474	3′	36 - 41	TTTATT
35	Hs.205785	AI754944	183	3′	36 - 41	TTTATT
36	Hs.206064	AI754534	444	3′		AATAAA(+)
37	Hs.228701	AA666138	315	3′	289-294	AATAAA(+)
38	Hs.230929	AI754032	458	3′	34-39	TTATT
39	Hs.236504	AI753241	392	3′		
40	Hs.236505	AI754092	509	3′		
41	Hs.236506	AI754163	462	3′		
42	Hs.236508	AI755079	474	3′	39-44	TTATT
43	Hs.241930	AW068972	441	3′		
44	Hs.241931	AW069084	477	3′		
45	Hs.241932	AW069546	301	3′		
46	Hs.243236	AW069016	514	3′		
47	Hs.243237	AW069415	335	3′	42-47	TAATTT

Table 1 continued on next page

TABLE 1: Continued

Clone number	Hs.ID	GenBank ID	EST read length (bp)	Seq. read direction	PolyA_Sig position	PolyA_Seq (-)
48	Hs.243238	AW069660	487	3′		
49	Hs.243997	AW069073	459	3′		
50	Hs.243998	AW069500	467	3′	34-39	TTTATT
51	Hs.243999	AW069787	436	3′	21-26	TTTATT
52	Hs.244767	AW069550	483	3′	474-479	AATAAA(+)
53	Hs.245546	AW068467	484	3′	42-47	TTTATT
54	Hs.246401	AW069657	468	3′		
55	Hs.246402	AW069832	453	3′	34-39	TTTAAT
56	Hs.246403	AW069867	460	3′	47-52	TTTAAT
57	Hs.257593	AA669840	420	3′	51-56	TTTAAT
58	Hs.259234	AW069762	297	3′	37-42	TTTAAT
59	Hs.204931	AW069679	615	3′	33-38	TTTATT

Novel ESTs found in HBMSC cDNA library with UniGene (Hs) ID and serial number, GenBank ID, length of EST read, read direction (3′), and position and sequence of the poly(A) site on these ESTs.

number, leading to decreased bone remodeling with a negative bone balance and causing profound osteopenia [33].

We assessed the cellular function of the gene products for 1030 known gene sets found in the HBMSC cDNA library based on the TIGR (The Institute for Genome Research, Rockville, Maryland) gene cellular function directory, which lists gene products according to the following functions: (1) cell division, (2) cell signaling and communication, (3) cell structure/motility, (4) cell/organism defense, (5) gene/protein synthesis, (6) metabolism, and (7) unclassified (Table 2). Among the 30 most abundant genes were 13 genes (43.3%) whose products serve a cell structure or motility function. There were also 13 genes (43.3%) encoding cell signaling or communication proteins. There were two genes (6.6%) related to gene/protein expression and two genes (6.6%) that belong to the unclassified gene group.

In addition to the highly expressed known genes in HBMSC, certain other EST clusters not related to known genes were also highly expressed. We refer to these as unknown EST clusters. Table 3 lists the 14 most highly expressed unknown EST clusters in HBMSC. The most highly expressed EST cluster in HBMSC (identified with UniGene cluster Hs. 40098) accounted for almost 0.6% of the total gene transcripts. The second most highly expressed EST clusters (UniGene Hs. 16869) accounted for 0.3% of total gene expression. Presumably, these ESTs represent genes that are critical for growth and development as well as cell specificity of HBMSC. Further characterization of these ESTs will give us more detailed information about their functions and roles. Among the unknown gene clusters are some with high homology, but not complete identity, to known genes. For example EST Hs. 196711 (UniGene ID) is highly similar to human extracellular protein, whereas another EST, Hs. 198089 (UniGene ID), is highly similar to human lysyl hydroxylase 2. These ESTs may represent new members of previously identified gene families (Table 3).

Table 4 (see supplemental data) lists all the functional categorized known gene transcripts of HBMSC obtained from the analysis of 4258 ESTs and sorted by cellular functional category. A total of 1030 distinct sequences were identified, each corresponding to a different transcript. Based on their cellular functions, we categorized 1030 distinct transcripts into seven groups (see above). The categorized gene transcripts were ordered according to the times (frequency) each transcript was found. Table 4 also shows the gene expression level, the gene symbol (if assigned), the gene location (if known), and the UniGene title of these known genes in HBMSC.

New Members of the Known Genes

A major application of high-throughput EST sequencing analysis is to explore families of related genes [28]. Through BLAST searching of NCBI's databases, we have found that certain of the less highly expressed ESTs from the HBMSC cDNA library also share high sequence similarity to known genes. Table 5 shows that at least 20 single ESTs in this cDNA library have high similarity to a variety of different known genes or gene families. For example, among the extracellular matrix proteins (cell structure or motility group), HBMSC EST clone 5131547 is highly similar to the EGF-containing fibulin-like extracellular protein (97% identity), whereas HBMSC EST clone 5132794 is highly similar to a protein expressed in fibroblasts of periodontal ligament (98% identity). Among the cell cycle progression proteins (cell division group), HBMSC EST clone 5132949 is highly similar to the cell cycle progression restoration-8 protein (93% identity). Among the ADP-ribosylation factor-like binding proteins (gene and protein expression group), HBMSC EST clone 5133421 is highly similar to the ARF-like-2 binding protein BART1 (98% identity). Among the hypothetical protein family (cell and organism defense group), clone 1027353 is highly similar to the yeast (*Saccharomyces cerevisiae*) hypothetical 54.2-kD

doi:10.1006/geno.2001.6683, available online at http://www.idealibrary.com on IDEAL

TABLE 2: Top 30 most highly expressed genes in HBMSC cDNA library

Rank	Stomal clones	Expression level	UniGene ID	Gene symbol	Chromosome location	UniGene title	Cellular function
1	188	4.65%	Hs.118162	*FN1*	2q34	fibronectin 1	2
2	90	2.23%	Hs.179573	*COL1A2*	7q22.1	collagen, type I, α 2	3
3	82	2.03%	Hs.172928	*COL1A1*	17q21.3-q22	collagen, type I, α 1	3
4	78	1.93%	Hs.111779	*SPARC*	5q31-q33	secreted protein, acidic, cysteine-rich (osteonectin)	3
5	74	1.83%	Hs.181165	*EEF1A1*	6q14	eukaryotic translation elongation factor 1 α 1	5
6	71	1.76%	Hs.215747	*ACTG1*	17q25	actin, γ 1	3
7	66	1.63%	Hs.180952	*ACTB*	7p15-p12	actin, β	3
8	44	1.09%	Hs.75777	*TAGLN*	11q23.2	transgelin	3
9	42	1.04%	Hs.62954	*FTH1*	11q13	ferritin, heavy polypeptide 1	2
10	41	1.02%	Hs.217493	*ANX2*	15q21-q22	annexin II (lipocortin II; calpactin I, heavy polypeptide)	2
11	31	0.77%	Hs.75511	*CTGF*	6q23.1	connective tissue growth factor	2
12	31	0.77%	Hs.118787	*TGFBI*	5q31	transforming growth factor, β-induced, 68 kD	3
13	29	0.72%	Hs.195188		2	human normal keratinocyte substraction library mRNA, clone H22a, complete sequence	7
14	28	0.69%	Hs.2064	*VIM*	10p13	vimentin	3
15	27	0.67%	Hs.169476	*K-ALPHA-1*	9	tubulin, α, ubiquitous	3
16	27	0.67%	Hs.195851		10	Human aortic-type smooth muscle α-actin (SM-α-A) gene, exon 9	3
17	25	0.62%	Hs.1516	*IGFBP4*	17q12-q21.1	insulin-like growth factor-binding protein 4	2
18	24	0.59%	Hs.82085	*PAI1*	7q21.3-q22	plasminogen activator inhibitor, type I	2
19	17	0.42%	Hs.73742	*RPLP0*	12	ribosomal protein, large, P0	5
20	16	0.40%	Hs.76152	*DCN*	12q23	decorin	2
21	15	0.37%	Hs.207396		15	*H. sapiens* Opa-interacting protein OIP3 mRNA, partial cds	2
22	15	0.37%	Hs.119571	*COL3A1*	2q31	collagen, type III, α 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)	3
23	13	0.32%	Hs.7753	*CALU*	7q32	calumenin	2
24	13	0.32%	Hs.179661		6	*H. sapiens* clone 24703 β-tubulin mRNA, complete cds	3
25	13	0.32%	Hs.5831	*TIMP1*	Xp11.3-p11.23	tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	2
26	12	0.30%	Hs.87409	*THBS1*	15q15	thrombospondin 1	2
27	12	0.30%	Hs.111334	*FTL*	19q13.3-q13.4	ferritin, light polypeptide	2
28	11	0.27%	Hs.82985	*COL5A2*	2q14-q32	collagen, type V, α 2	3
29	11	0.27%	Hs.119122			*H. sapiens* mRNA for 23 kD highly basic protein	7
30	11	0.27%	Hs.74487		10	human β-1D integrin mRNA, cytoplasmic domain, partial cds	2

Columns 1–3 indicate gene rank, clone number, and expression level. Columns 4–7 indicate UniGene ID, chromosome location, and name. Column 8 indicates cellular function (see text).

TABLE 3: Top 14 most highly expressed EST clusters in HBMSC cDNA library

Rank	Stomal clones	Expression level	UniGene ID	Chomosome location	UniGene title
1	24	0.59%	Hs.40098	15	ESTs
2	12	0.30%	Hs.16869	6	ESTs
3	11	0.27%	Hs.155712	3	ESTs
4	9	0.22%	Hs.41271	3	ESTs
5	9	0.22%	Hs.216036	15	ESTs
6	8	0.20%	Hs.25035	1	ESTs
7	7	0.17%	Hs.14838	1	ESTs
8	6	0.15%	Hs.87428	1	ESTs
9	6	0.15%	Hs.9315	1	ESTs, weakly similar to pancortin-1 [*M. musculus*]
10	5	0.12%	Hs.8687	5	ESTs
11	5	0.12%	Hs.70823	8	ESTs
12	5	0.12%	Hs.196711	2	ESTs, highly similar to extracellular protein [*H. sapiens*]
13	4	0.10%	Hs.30343	3	ESTs
14	4	0.10%	Hs.198089	3	ESTs, highly similar to lysyl hydroxylase isoform 2 [*H. sapiens*]

Columns 1–3 indicate rank, cluster number, and expression level. Columns 4–6 indicate UniGene ID, chromosome location, and UniGene title. Three EST clusters share similarities to other known genes to different degrees (high similarity means two DNA sequences with ≥ 90% identity; weakly similar means two DNA sequences with < 70% identity; moderately similar is in between).

protein in the CDC12-ORC6 intergenic region, whereas EST clones 1119758 and 1119800 are highly similar to the hypothetical 67.6-kD protein ZK637.3 (*Caenorhabditis elegans*) and the hypothetical 43.2-kD protein C34E10.13 (*C. elegans*), respectively. Further analysis of these EST clones will facilitate recognition of the expression and function of the gene products represented by these ESTs, because many of these known genes have already been explored to define the pattern of gene expression, protein biochemistry, and cellular function of their products in human cells or other organisms.

Digital Gene Localization

Both STSs and ESTs can be used for mapping gene locations [28]. To identify the gene location of the HBMSC ESTs, we did BLAST searches against both NCBI's STS database and the high-throughput genomic sequence database (HTGS). This analysis, which we call digital gene localization (DGL), can determine the precise chromosomal location of an EST if an exact match is made to a genomic sequence. We used the NCBI network BLAST analysis with high-stringency parameters (> 100 bp long, 98% identity) for DGL. So far, from this analysis, 34 unknown ESTs from the HBMSC cDNA library, comprising 34 different genes, have been assigned to chromosomal locations. Table 6 shows that these ESTs have been located on 14 of the 23 human chromosome pairs. The DGL analysis also predicts the gene structure, including intron/exon junctions, of these genes.

We also analyzed genes and ESTs from HBMSC for their chromosome location distribution. Table 7 shows all the genes and ESTs with known chromosome location information from the HBMSC cDNA library compared with the whole UniGene mapping data. This analysis shows that the gene distribution from HBMSC is quite similar to the UniGene mapping distribution.

Discussion

A rapidly accelerating amount of information on expressed gene sequences has been generated in the past few years by researchers in the genome community. As of October 30, 2001, there are 6954 human cDNA libraries from 195 different organs, tissues, cells, and cell lines in the NCBI UniGene database. Partial and complete sequences from clones in these libraries have been combined with other information in GenBank and dbEST to form the UniGene collection of > 96,332 cDNA clusters, representing about 96,332 gene transcripts.

The high-throughput single-pass partial sequencing of cDNAs to generate ESTs has proven to be a powerful and successful way to assemble a profile of genes expressed in a particular organism, tissue, or cell type [18–20,22–24,34]. Libraries of short cDNA fragments corresponding to the 3′ or 5′ end regions of the mRNA have offered advantages to accelerate gene discovery and gene mapping.

The use of ESTs to produce transcript maps provides a significant aid to positional cloning of genes involved in human genetic diseases [21]. The ESTs generated in this study currently are being used to create such a transcript map for HBMSC that will provide a framework for identifying genes involved in important skeletal and hematopoietic phenotypes. Comparative EST analysis also provides a means for identifying polymorphisms that may also be useful for genetic mapping studies. HBMSC are pluripotent cells with the potential to differentiate into different cell types, including osteoblasts, chondrocytes, myelosupportive stroma, and adipocytes. Our data serve as an important resource to facilitate further study of the relationship between gene expression and differentiated phenotypes of HBMSC.

In practical terms, the work reported here should benefit the study of HBMSC in several ways. First, for the ESTs that are found only in this cDNA library, it will be of interest to

doi:10.1006/geno.2001.6683, available online at http://www.idealibrary.com on IDEAL

TABLE 5: HBMSC ESTs which are highly similar to the known genes

No.	HBMSC Clone	GID	Similar to the known genes' UniGene title	Score	*E*-value	Identities	Gaps
1	HBMSC_cr08a05	5131547	EST, Highly similar to EGF-containing fibulin-like extracellular protein	636	0	337/345, 97%	
2	IMAGE:1119354	2849166	EST, Highly similar to EGF-containing fibulin-like extracellular protein	950	0	510/518, 98%	3/518
3	HBMSC_cr26e06	5132794	EST, Highly similar to expressed in fibroblasts of periodontal igament	894	0	484/493, 98%	3/493
4	IMAGE:1026779	2631738	EST, Highly similar to expressed in fibroblasts of periodontal igament	605	e-171	334/340, 98%	4/310
5	HBMSC_cr28g05	5132949	EST, Highly similar to cell cycle progression restoration 8 protein	827	0	542/578, 93%	33/578
6	IMAGE:1119320	2848829	EST, Highly similar to *H. sapiens* clone 24761 mRNA sequence	1374	0	731/740, 98%	4/740
7	HBMSC_cr14c01	5131984	EST, Highly similar to *H. sapiens* clone 25022 mRNA sequence	995	0	559/579, 96%	3/579
8	HBMSC_cr36c04	5133421	EST, Highly similar to Arf-like 2 binding protein BART1	272	e-151	288/291, 98%	2/291
9	IMAGE:1091551	2432964	EST, Highly similar to Arf-like 2 binding protein BART1	182	8.00E-98	199/202, 98%	2/202
10	IMAGE:1026726	2618436	EST, Highly similar to NG-dimethylarginine dimethylaminohydrolase homolog mRNA	801	0	427/432, 98%	2/432
11	HBMSC_cr04c02	5131307	EST, Highly similar to *Oryctolagus cuniculus* ubiquitin-conjugating enzyme E2-32K	351	1.00E-94	216/229, 94%	
12	HBMSC_cr27e07	5132865	EST, Highly similar to *O. cuniculus* ubiquitin-conjugating enzyme E2-32K	341	1.00E-91	218/232, 93%	1/232
13	IMAGE:1119518	2631410	EST, Highly similar to citb_175_g_20	603	e-170	320/324, 98%	1/324
14	IMAGE:1119521	2631416	EST, Highly similar to citb_175_g_20	565	e-159	306/314, 97%	1/314
15	IMAGE:1091310	2433389	EST, Highly similar to protein kinase C and casein kinase substrate in neurons 2 (PACSIN2)	567	e-160	340/363, 93%	1/363
16	IMAGE:1091498	2432896	EST, Highly similar to *H. sapiens* CGI-06 protin	646	0	342/346, 98%	1/346
17	IMAGE:1027289	2433866	EST, Highly similar to *H. sapiens* sirtuin type 2 (SIRT2)	769	0	423/432, 97%	3/432
18	HBMSC_cr05b10	5131369	EST, Highly similar to surface 4 integral membrane protein	660	0	364/374, 97%	2/374
19	IMAGE:1071195	2433109	EST, Highly similar to *H. sapiens* homolog of *D. melanogaster* flightless-I gene product	874	0	528/548, 96%	8/548
20	HBMSC_cr15g05	5132087	EST, Highly similar to growth-factor inducible immediate early gene product CYR61	577	e-163	324/334, 97%	4/334
21	IMAGE:1119029	2631568	EST, Highly similar to mouse ubiquitin-conjugating enzyme UbcM2	581	e-164	347/362, 95%	11/362
22	HBMSC_cr36b06	5133416	EST, Highly similar to human APMCF1	489	0	522/528, 98%	4/528
23	IMAGE:1119985	2714066	EST, Highly similar to mouse protein B gene	172	4.00E-41	112/119, 94%	1/119
24	HBMSC_cr26f02	5132802	EST, Highly similar to human host cell factor homolog LCP mRNA	954	0	500/506, 98%	1/506
25	HBMSC_cr29c08	5132989	EST, Highly similar to secretory carrier membrane protein 3 [*H. sapiens*]	882	0	458/464, 98%	
26	IMAGE:1119154	2631509	EST, Highly similar to Fn54 mRNA [*M. musculus*]	204	6.00E+51	153/170, 90%	4/170
27	IMAGE:1119433	2631363	EST, Highly similar to KIAA0095 gene is related to S. cerevisiae NIC96 gene [*H. sapiens*]	720	0	421/439, 95%	1/439
28	IMAGE:1091233	2433451	EST, Highly similar to sorting nexin 2 [*H. sapiens*]	456	e-126	264/270, 97%	4/270
29	HBMSF1F8	2307058	EST, Highly similar to protein phosphatase 4, regulatory subunit 1 [*H. sapiens*]	589	e-166	311/315, 98%	1/315

Columns 1 and 2 indicate EST clone name and GenBank ID. Column 3 indicates UniGene title of known genes that are highly similar to ESTs found in the HBMSC cDNA library. Columns 4–7 indicate score, *E*-value, identities (%), and gaps.

TABLE 6: Digital gene localization of HBMSC ESTs

HBMSC		STS			*E*-	HTGS		Score	*E*-	Chromosome
clone_ID	GI	hit_GI	Description	Score	Value	hit_GI	Description		Value	location
HBMSC_cr02d02	5131188	1347452	human STS EST221972	577	e-164	5441631	Human chromosome 6 clone 349A12	730	0.0	6p21
HBMSC_cr03c09	5131247	860053	human STS WI-8010	775	0.0	5306297	*H. sapiens* clone 44_C_14,	842	0.0	9q32
HBMSC_cr04a02	5131288	1017526	human STS SHGC-11260	513	e-145	4662688	*H. sapiens* clone DJ0042M02	513	e-144	7p22
HBMSC_cr06c12	5131436	1593001	human STS SHGC-33702	630	e-180	5441433	Human chromosome 6 clone J238D15	638	0.0	6q12
HBMSC_cr07a02	5131474	1593001	human STS SHGC-33702	624	e-179	5441433	Human chromosome 6 clone J238D15	632	e-179	6q14
HBMSC_cr08a06	5131548	1348911	human STS STS_D11566	680	0.0	5457183	Human chromosome X clone A123M24	904	0.0	1p11
HBMSC_cr10b02	5131680	1161760	human STS CHLC.UTR_	194	2e-49	5230407	Human chromosome 7 clone RP11-754B14	194	2e-47	7
HBMSC_cr10g12	5131715	1342161	human STS WI-14149	537	e-152	4160138	Human clone pDJ416j11	680	0.0	11p15
HBMSC_cr13e04	5131924	1375158	human STS SHGC-31592	359	1e-98	5174845	*H. sapiens* clone 44_A_12,	868	0.0	2p14
HBMSC_cr14f03	5132011	1348373	human STS EST65244	936	0.0	4995278	Human chromosome 6 clone 340B19	952	0.0	6p21
HBMSC_cr15b09	5132045	1340963	human STS A005Z27	355	1e-97	5523805	*H. sapiens* clone NH0471A05	686	0.0	2q31
HBMSC_cr16b04	5132111	1593924	human STS SHGC-36880	648	0.0	3213020	*H. sapiens* clone DJ1152C17	769	0.0	7q31
HBMSC_cr16e05	5132135	4192174	WIAF-1830-STS	246	1e-64	4895146	Human chromosome 12 clone 917O5	827	0.0	12
HBMSC_cr19g12	5132371	1131946	human STS SHGC-15798	383	e-106	5523787	*H. sapiens* clone RPCI11-412D9	751	0.0	12q24
HBMSC_cr20e09	5132401	1593001	human STS SHGC-33702	620	e-177	5441433	Human chromosome 6 clone J238D15	628	e-178	6q13
HBMSC_cr23a03	5132556	1347452	human STS EST221972	583	e-166	5441631	Human chromosome 6 clone 349A12	720	0.0	6p22
HBMSC_cr29e12	5133010	1340966	human STS A005Z33	214	4e-55	3334548	Human DNA sequence clone 316D5	484	e-135	22q13
HBMSC_cr30g05	5133089	1592931	human STS SHGC-32576	547	e-155	5102597	Human sapiens clone R-280K24	987	0.0	14
HBMSC_cr35b10	5133344	1340940	human STS A005Y34	266	e-70	3212909	*H. sapiens* clone RG271G1	860	0.0	7p15
HBMSC_cr37g04	5133533	1347452	human STS EST221972	583	e-166	5441631	Human chromosome 6 clone 349A12	884	0.0	6p21
IMAGE:1026737	2618450	1375158	human STS SHGC-31592	355	1e-97	5174845	*H. sapiens* clone 44_A_12	880	0.0	2p14
IMAGE:1027308	2433871	1341334	human STS WI-6601	373	e-103	4585946	*H. sapiens* clone hCIT.58_E_17	692	0.0	17
IMAGE:1027332	2433892	1341334	human STS WI-6601	373	e-103	4585946	*H. sapiens* clone hCIT.58_E_17	728	0.0	17
IMAGE:1071206	2433103	860053	human STS WI-8010	771	0.0	5306297	*H. sapiens* clone 44_C_14	918	0.0	9q32
IMAGE:1090505	2433556	1594096	human STS SHGC-37377	505	e-143	3057011	*H. sapiens* clone pDJ460g16	837	0.0	15q26
IMAGE:1090602	2432990	1343190	human STS WI-17110	505	e-143	4895146	Human chromosome 12 clone 917O5	537	e-151	12q22
IMAGE:1090712	2433159	1340524	human STS A002D16	329	8e-90	3334990	Human chromosome 4, clone B286M7	490	e-136	4
IMAGE:1091310	2433389	1340966	human STS A005Z33	214	3e-55	3334548	Human DNA from clone 316D5	567	e-160	Xq25
IMAGE:1091436	2432841	2996700	Human STS SHGC-56773	730	0.0	3242690	*H. sapiens* clone C0164F16	920	0.0	4
IMAGE:1119199	2631542	1593874	human STS SHGC-36771	585	e-167	3212928	*H. sapiens* clone RG099B05	624	e-177	10q22
IMAGE:1119265	2849073	860184	human STS WI-8550	599	e-171	4826437	Human chromosome 1 clone 120G22	618	e-175	1
IMAGE:1119530	2620406	1396296	human STS SHGC-32527	609	e-174	5001498	*H. sapiens* clone NH0420C09	730	0.0	2p13
IMAGE:1119881	2713906	1344208	human STS WI-30085	505	e-142	5508868	Human chromosome 11 clone CTC-366J2	922	0.0	11q
IMAGE:1119912	2713978	1244202	human STS SHGC-11975	478	e-134	4680453	*H. sapiens* clone NH0310K15	858	0.0	2q34

A search of HTGS and STS databases reveals 34 ESTs from the HBMSC cDNA library that have DNA sequences on different chromosomes. Columns 1 and 2 indicate EST clone ID and GenBank ID. Columns 3 and 4 indicate STS hit ID and description in the STS database. Columns 5 and 6 indicate hit score and *E*-value. Columns 7–10 indicate HTGS hit ID, description, hit score, and *E*-value. Column 11 indicates chromosome location of these ESTs.

doi:10.1006/geno.2001.6683, available online at http://www.idealibrary.com on IDEAL

TABLE 7: Gene distribution of HBMSC on the human chromosomes based on UniGene build #133, released April 20, 2001

Chromosome	No. of UniGene clusters of stromal	No. of UniGene clusters of all	% of mapped UniGene stroma	% of mapped UniGene all
1	199	2031	11.06%	9.93%
2	129	1492	7.17%	7.29%
3	142	1298	7.89%	6.34%
4	57	918	3.17%	4.49%
5	104	1005	5.78%	4.91%
6	92	1166	5.11%	5.70%
7	76	1037	4.22%	5.07%
8	67	784	3.72%	3.83%
9	66	865	3.67%	4.23%
10	81	870	4.50%	4.25%
11	112	1179	6.23%	5.76%
12	105	1105	5.84%	5.40%
13	44	431	2.45%	2.11%
14	59	683	3.28%	3.34%
15	68	699	3.78%	3.42%
16	47	660	2.61%	3.23%
17	80	994	4.45%	4.86%
18	24	354	1.33%	1.73%
19	81	986	4.50%	4.82%
20	46	479	2.56%	2.34%
21	20	246	1.11%	1.20%
22	42	470	2.33%	2.30%
X	57	686	3.17%	3.35%
Y	1	24	0.06%	0.12%
			100.00%	100.00%

All genes and ESTs with a known chromosome location from the HBMSC cDNA library compared with the whole UniGene mapping data. This analysis shows that gene distribution from HBMSC is quite similar to the UniGene mapping distribution.

determine their biological functions in the growth and development of HBMSC. Second, the ESTs from HBMSC can be used to develop STSs and can be definitively mapped. Third, for those interesting gene transcripts, such as new members of known gene families, ESTs from HBMSC can be used to obtain full-length cDNA clones by library screening or 5′ rapid amplification of cDNA ends [35]. Fourth, the gene expression profile of HBMSC can be used as a reference for comparative gene expression pattern studies with differentiated HBMSC and skeletal-related cells. Fifth, by using gene cluster information, a bone-enhanced cDNA microarray can be developed and used to study gene expression in skeletal tissue at different stages of growth and development in health and disease.

Although generation of ESTs and data file analysis are the first steps to further understanding the gene expression and cellular phenotype of HBMSC, the reagents and data reported here can provide important and useful information for the skeletal research community. All sequenced EST clones from the HBMSC cDNA library are already available to the public. Researchers interested in any of these EST clones may obtain them by contacting us or through Research Genetics. An SGAP web site, which includes bone cDNA library information and data analysis as well as a bone-related gene database, is available at http://sgd.nia.nih.gov/. In conjunction with genome-wide EST mapping projects [36] and CGAP [16] as well as genomic sequencing, our studies should accelerate the process of gene discovery and functional genomic analysis of skeletal growth and development in health and disease and enable a greater understanding of the pathophysiology of skeletal disorders.

Supplementary data for this article are available on IDEAL (http://www.idealibrary.com).

Materials and Methods

Construction of HBMSC cDNA library. To maximize the gene representation of the cDNA library for purposes of gene discovery and gene expression profiling, we selected mixed cells derived from three donors (32-year-old black female, 35-year-old black male, and 43-year-old white male). Cell lines from HBMSC derived from normal volunteer donors under Institutional Review Board (IRB) approved guidelines (94-D-0186) were established according to a previously published method [9]. HBMSC preparations from aspirate and surgical specimens were passed consecutively through 16- and 20-gauge needles to break up cell aggregates for obtaining single cell suspensions. After primary culture, HBMSC cultures with large numbers of colonies were combined. Later passages were performed when cells were approaching confluence. For RNA isolation, we used multicolony-derived HBMSC strains at the second or third passage. The cDNA library was constructed in lambda ZapII (Stratagene). Extracted total RNA was isolated from primary cultures of HBMSC by the CsCl gradient centrifugation procedure [37], and poly(A)$^+$ mRNA was obtained by affinity chromatography on an oligo(dT)-cellulose column (5 Primer & 3 Primer, Inc.). We used about 10 μg poly(A)$^+$ to construct a lambda ZapII library (custom library section of Stratagene Cloning Systems). Double-stranded cDNA was cloned into *Eco*RI/*Xho*I restriction sites of lambda ZapII. pBluescript SK+ phagemids were obtained by en masse *in vivo* excision of lambda Zap clones [38] by coinfecting *E. coli* XL-1Blue cells with the ExAssist helper phage (Stratagene). The excised phagemids were used to infect *E. coli* SOLR cells (Stratagene) for production of double-stranded DNA templates. Transformants were plated onto LB agar containing ampicillin (100 μg/mL).

DNA sequencing. We isolated plasmid DNA for sequencing in a 96-well configuration as described elsewhere (http://genome.wustl.edu/gsc/Protocols/pucprep.shtml). Fluorescent sequencing was done with one-quarter strength BigDye terminator chemistry (Perkin Elmer/Applied Biosystems, Foster City, CA) and Tetrad thermal cyclers (MJ Research, Waltham, MA) according to the manufacturer's recommendations. Sequencing reactions were analyzed on ABI Prism 377xl automated DNA sequencing instruments (Perkin Elmer/Applied Biosystems, Foster City, CA). The 3′ end of cDNA clones were sequenced with the universal -21M13 forward primer (5′-TGTAAAAC-GACGGCCAGT-3′). The 5′ end of cDNA clones were sequenced with the universal M13RP1 reverse primer (5′-CAGGAAACAGCTATGACC-3′).

DNA sequence data analysis. A total of 4258 EST sequences from the HBMSC cDNA library were analyzed; 2550 EST sequences were sequenced by NISC, and 1708 EST sequences were sequenced at Washington University as part of the Merck/Washington University EST sequencing project [28]. The sequencing data from Merck/Washington University were extracted from daily dbEST FASTA (Fast All, a computer format) updates. The sequences were inserted into a relational database. Automated processes were developed and used for various sequence analyses. Using the BLAST network client to access BLAST [39] at the NCBI, we obtained BLASTN (BLAST nucleotide) results for each sequence against both the nonredundant nucleotide and dbEST databases. The homology results were parsed using the BTAB (BLAST Tabulator) program and inserted into the relational database. Cross-reference tables for mapping all dbEST sequence identifiers to clone IDs and library IDs and mapping of sequence identifiers to UniGene clusters are maintained in the database. This allows simple SQL (structure quarry language) procedures to generate profiles of clones based on libraries of hits to other dbEST sequences as well as UniGene clustering [40].

Acknowledgments

We thank Gerard Bouffard, Jeffrey Touchman, and the staff of NISC for DNA sequencing and analysis. We also thank Greg Lennon from Lawrence Livermore National Laboratory, Berkeley (now at Gene Logic); Robert Waterston and Richard Wilson from Washington University School of Medicine, St. Louis, for part of the EST sequencing (as part of the Merck-EST project); Eric Green of NHGRI; David Krizman of NCI; and Greg Schuler and Mark Boguski of NCBI for thoughtful discussion. We are grateful to OrthoGene of New York, which provided partial support for EST sequencing through a Cooperative Research and Development Agreement (CRADA) with NHGRI/NIH

**RECEIVED FOR PUBLICATION JUNE 20;
ACCEPTED NOVEMBER 2, 2001.**

References

1. Prockop, D. J. (1997). Marrow stromal cells as stem cells for nonhematopoietic tissues. *Science* **276:** 71–74.
2. Marx, J. C., *et al.* (1999). High-efficiency transduction and long-term gene expression with a murine stem cell retroviral vector encoding the green fluorescent protein in human marrow stromal cells. *Hum. Gene Ther.* **10:** 1163–1173.
3. Chiang, G. G., *et al.* (1999). Bone marrow stromal cell-mediated gene therapy for hemophilia A: in vitro expression of human factor VIII with high biological activity requires the inclusion of the proteolytic site at amino acid 1648. *Hum. Gene Ther.* **10:** 61–76.
4. Prockop, D. J. (1998). Marrow stromal cells as stem cells for continual renewal of nonhematopoietic tissues and as potential vectors for gene therapy. *J. Cell Biochem. Suppl.* **30-31:** 284–285.
5. Friedenstein, A. J. (1995). Marrow stromal fibroblasts. *Calcif. Tiss. Int.* **56** (**suppl.1**)**:** S17.
6. Friedenstein, A. J., *et al.* (1974). Stromal cells responsible for transferring the microenvironment of the hemopoietic tissues, cloning in vitro and retransplantation in vivo. *Transplantation* **17:** 331–340.
7. Owen, M., and Friedenstein, A. J. (1988). Stromal stem cells: marrow-derived osteogenic precursors. *Ciba Found. Symp.* **136:** 42–60.
8. Kuznetsov, S. P., and Robey, P. G. (1996). Species differences in growth requirements for bone marrow stromal fibroblast colony formation *in vitro. Calcif. Tiss. Int.* **59:** 265–270.
9. Kuznetsov, S. A., Friedenstein, A. J., and Robey, P. G. (1997). Factors required for bone marrow stromal fibroblast colony formation *in vitro. Br. J. Haematol.* **97:** 561–570.
10. Simmons, P. J., and Torok-Storb, B. (1991). Identification of stromal cell precursors in human bone marrow by a novel monoclonal antibody, STRO-1. *Blood* **78:** 55–62.
11. Stewart, K., *et al.* (1999). Further characterization of cells expressing STRO-1 in cultures of adult human bone marrow stromal cells. *J. Bone Miner. Res.* **14:** 1345–1356.
12. Gordon, M. Y., Lewis, J. L., Marley, S. B., Grand, F. H., and Goldman, J. M. (1997). Stromal cells negatively regulate primitive haemopoietic progenitor cell activation via a phosphatidylinositol-anchored cell adhesion/signalling mechanism. *Br. J. Haematol.* **96:** 647–653.
13. Morel, F., *et al.* (1998). Equal distribution of competitive long-term repopulating stem cells in the CD34$^+$ and CD34$^-$ fractions of Thy-1 low Lin$^-$/Sca-1+ bone marrow cells. *Exp. Hematol.* **26:** 440–448.
14. Choi, S. J., *et al.* (1998). Cloning and identification of human Sca as a novel inhibitor of osteoclast formation and bone resorption. *J. Clin. Invest.* **102:** 1360–1368.
15. Jia, L., *et al.* (1997). SGAP: The skeletal genome anatomy project. *Am. J. Hum. Genet.* **61:** A378.
16. Strausberg, R. L., Dahl, C. A., and Klausner, R. D. (1997). New opportunities for uncovering the molecular basis of cancer. *Nat. Genet.* **Special:** 415–416.
17. Ho, N. C., Jia, L. B., Driscoll, C. C., Gutter, E. M., and Francomano, C. A. (2000). A skeletal gene database. *J. Bone Miner. Res.* **15:** 2095–2122.
18. Adams, M. D., *et al.* (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.
19. Adams, M. D., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4:** 256–267.
20. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4:** 373–380.
21. Adams, M. D., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.
22. Khan, A. S., *et al.* (1992). Single pass sequencing and physical and genetic mapping of human brain DNAs. *Nat. Genet.* **2:** 180–185.
23. McCombie, W. R., *et al.* (1992). *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nat. Genet.* **1:** 124–131.
24. Okubo, K., *et al.* (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2:** 173–179.
25. Matsubara, K., and Okubo, K. (1993). Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotechnol.* **4:** 672–677.
26. Hillier, L., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.
27. Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST—database for "expressed sequence tags." *Nat. Genet.* **4:** 332–333.
28. Gerhold, D., and Caskey, T. (1996). It's the genes! EST access to human genome content. *Bioessays* **18:** 973–981.
29. Milner, R. J., and Sutcliffe, J. G. (1983). Gene expression in rat brain. *Nucleic Acids Res.* **11:** 5497–5520.
30. Putney, S. D., Herlihy, W. C., and Schimmel, P. (1983). A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302:** 718–721.
31. Ajioka, J. W., *et al.* (1998). Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the apicomplexa. *Genome Res.* **8:** 18–28.
32. Touchman, J. W., *et al.* (1997). 2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries. *Genome Res.* **7:** 281–292.
33. Delany, A. M., *et al.* (2000). Osteopenia and decreased bone formation in osteonectin-deficient mice. *J. Clin. Invest.* **105:** 915–923.
34. Waterston, R., *et al.* (1992). A survey of expressed genes in *Caenorhabditis elegans*. *Nat. Genet.* **1:** 114–123.
35. Frohman, M. A., Dush, M. K., and Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide. *Proc. Natl. Acad. Sci. USA* **85:** 8998–9002.
36. Schuler, G. D., *et al.* (1997). A gene map of the human genome. *Science* **274:** 540–546.
37. Ibaraki, K., Termine, J. D., Whitson, S. W, and Young, M. F. (1992). Bone matrix mRNA expression in differentiating fetal bovine osteoblasts. *J. Bone Miner. Res.* **7:** 743–754.
38. Alting-Mees, M., *et al.* (1992). New lambda and phagemid vectors for prokaryotic and eukaryotic expression. *Strategies* **5:** 58–61.
39. Altschul, S. F., *et al.* (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
40. Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75:** 694–698.